Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants

Will Fischer[1,7], Simon Perkins[1,7], James Theiler[1], Tanmoy Bhattacharya[1,2], Karina Yusim[1], Robert Funkhouser[1], Carla Kuiken[1], Barton Haynes[3], Norman L. Letvin[4], Bruce D. Walker[5], Beatrice H. Hahn[6], Bette T. Korber[1,2]

[1]Los Alamos National Laboratory, Los Alamos, NM 87545 USA

[2]Santa Fe Institute, Santa Fe, NM 87544 USA

[3]Duke University School of Medicine, Durham, NC 27710 USA

[4]Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02215 USA

[5]Infectious Disease Division, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02129, USA

[6]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[7]These two authors contributed equally to this study

Correspondence should be addressed to: Bette Korber: btk@lanl.gov

Abstract

HIV-1/AIDS vaccines must address the extreme diversity of HIV-1. We have designed novel polyvalent vaccine antigens comprised of sets of "mosaic" proteins, assembled from fragments of natural sequences via a computational optimization method. Mosaic proteins resemble natural proteins, and a mosaic set maximizes the coverage of potential T-cell epitopes (nine-amino-acid peptides) for a viral population. Coverage of viral diversity using mosaics is greatly increased compared to natural-sequence vaccine candidates for both variable and conserved proteins; for conserved HIV-1 proteins, global coverage may be feasible. For example, four mosaic proteins perfectly (9/9) match 74% of potential epitopes in global Gag sequences, and partially (8/9) match 87%, while a single natural Gag protein covers only 37% (9/9) and 67% (8/9). Mosaics provide diversity coverage comparable to thousands of separate peptides, but since the fragments of natural proteins are compressed into a small number of native-like proteins, they are tractable for vaccines.

Introduction

Designing an effective HIV vaccine is a many-faceted challenge. The vaccine would preferably elicit an immune response capable of preventing infection, or, minimally, controlling viral replication, even though immune responses to natural infection fail to eliminate the virus[1] or protect from superinfection[2]. Potent vaccines optimized vectors, immunization protocols, and adjuvants[1], combined with antigens that can stimulate responses that cross-react against the diverse spectrum of circulating viruses[3,4]. The problems of influenza vaccinology highlight the challenge of HIV-1: human influenza strains diverge by 1-2% per year, yet vaccines often fail to elicit protection from one year to the next, necessitating frequent vaccine updates[4]. In contrast, co-circulating HIV-1 strains differ from one another by 20% or more in relatively conserved proteins, and up to 35% in the Envelope (Env) protein[3,4].

Different degrees of viral diversity in regional HIV-1 epidemics provide a potential hierarchy for vaccine design strategies. Some geographic regions recapitulate global diversity, with most known HIV-1 subtypes, or clades, co-circulating; others are dominated by two subtypes and their recombinants, still others by a single subtype. Even vaccines for single subtypes must address extensive within-clade diversity[5]; but, as international travel erodes geographic distinctions, all countries would benefit from a global vaccine.

We present the design of polyvalent vaccine antigen sets, focused on T lymphocyte responses, and optimized for either the common B and C subtypes or for all HIV-1 variants in global circulation [the HIV-1 Main (M) group]. Cytotoxic T-lymphocytes (CTL) directly kill infected, virus-producing host cells, recognizing them via viral protein fragments (epitopes) presented on infected cell surfaces by human leukocyte antigen (HLA) class I molecules. Helper T-cells orchestrate immune responses via cytokine release upon recognition of viral epitopes presented in the context of class II molecules. These two cell types often overlap in function; both will likely be crucial for an HIV-1 vaccine. CTL responses have been implicated in slowing disease progression[6]; vaccine-elicited cellular immune responses in nonhuman primates help control pathogenic SIV or SHIV, reducing the likelihood of disease after challenge[7]; and experimental depletion of $CD8^+$ T-cells results in increased viremia in SIV infected rhesus macaques[8]. Furthermore, CTL escape mutations are associated with disease progression[9], so vaccine-stimulated memory responses that block potential escape routes may be valuable.

The highly variable Env protein is the primary target for neutralizing antibodies against HIV, and a vaccine will likely require Env vaccine antigens optimized for antibody responses[10]. T-cell-directed vaccine antigens, in contrast, can target the more conserved proteins — but even the most conserved HIV-1 proteins are diverse enough that variation is an issue. Artificial central-sequence vaccine approaches (e.g., consensus sequences, in which every amino acid is found in a plurality of sequences, or maximum likelihood reconstructions of ancestral sequences[3,11-13]) are promising; nevertheless, even centralized strains provide limited coverage of HIV-1 variants, and consensus-based reagents fail to detect many autologous T-cell responses[14].

Single amino acid changes can allow an epitope to escape T-cell surveillance; since many T-cell epitopes differ between HIV-1 strains at one or more positions, potential responses to any single vaccine antigen are limited. Including multiple variants in a polyvalent vaccine could enable responses to a broader range of circulating variants, and could prime the immune system against common escape mutants[15]. Escape from one T-cell receptor may create a variant susceptible to another[16,17], so polyclonal responses to epitope variants may be beneficial[18].

Escape mutations that inhibit processing[19] or HLA binding[20] cannot be directly countered by T-cells with different specificities, but responses to overlapping epitopes may block even some of these escape routes.

We propose using a polyvalent vaccine comprising several “mosaic” proteins (or genes encoding these proteins). The candidate vaccine antigens are cocktails containing a small number of composite proteins, optimized to include the maximum number of potential T-cell epitopes from a set of viral proteins. The mosaics are generated from natural sequences: they resemble natural proteins, but systematically include common (and exclude rare) potential epitopes. Since $CD8^+$ epitopes are contiguous and typically nine amino-acids long, we generate and score sets of mosaics based on “coverage” of nonamers (stretches of nine contiguous amino acids , hereafter “nine-mers”) in the natural sequences (fragments of similar lengths are also well represented). This strategy provides the level of diversity coverage achieved by a massively polyvalent multiple-peptide vaccine, but with important advantages: it allows antigen delivery as intact proteins (or genes), which are likely to be processed as in natural infections, and excludes low-frequency epitopes that are irrelevant to circulating strains.

Results

Protein Variation. In conserved HIV-1 proteins most positions are essentially invariant, and most variable positions have only two to three amino acids occurring at appreciable frequencies, and variable positions are usually dispersed between conserved positions. Therefore, within the boundaries of a $CD8^+$ T-cell epitope (8-12 amino acids, typically nine), most population diversity can be covered with a few variants. We computed upper bounds for population nine-mer coverage of Gag, Nef, and Env (gp120) as the number of variants is increased (**Fig. 1**). In conserved regions (Gag P24, central Nef), 2-4 variants yield high population coverage. By contrast, in variable regions (e.g. gp120, Nef termini), even eight variants achieve only limited coverage. Since each new addition is rarer, the benefits of further additions diminish as the number of variants increases.

Vaccine design optimization strategies. Figure 1 shows idealized nine-mer coverage. In reality, high-frequency nine-mers often conflict — because of local covariation, the optimal amino acid for one nine-mer may differ from that for an overlapping nine-mer, so the relative benefits of each amino acid must be evaluated in combination with nearby variants. For example, Alanine (Ala) and Glutamate (Glu) might each frequently occur in adjacent positions, but if the Ala-Glu combination is rarely found in nature, it should be excluded from the vaccine. We investigated several optimization strategies: a greedy algorithm, a semi-automated compatible-nine-mer strategy, an alignment-based genetic algorithm (GA), and an alignment-independent GA.

The alignment-independent GA generated mosaics with the best population coverage. This GA generates a user-specified number of mosaic sequences from a set of unaligned protein sequences, explicitly excluding rare or unnatural epitope-length fragments (potentially introduced at recombination breakpoints) that could induce non-protective responses specific to the vaccine antigens. These candidate vaccine antigens resemble natural proteins, but are assembled from frequency-weighted fragments of database sequences recombined at homologous breakpoints (**Fig. 2** and Methods); they approach maximal nine-mer coverage of the input population.

Selecting HIV protein regions for an initial mosaic vaccine. For our initial design, we focused on protein regions meeting specific criteria: i) relatively low variability, ii) high levels of immune recognition in natural infection, iii) a high density of known epitopes and iv) either early responses upon infection *or* $CD8^+$ T-cell responses associated with good outcomes in infected patients. First we assessed the level of nine-mer coverage achievable for different HIV proteins (**Fig. 3**), generating a set of four mosaics for each protein using either the M group or the B- and C-subtypes alone, and scoring coverage on the C subtype sequences. Several results are notable: i) within-subtype optimization provides excellent within-subtype coverage, but substantially poorer between-subtype coverage ; ii) Pol and Gag have the most potential to elicit broadly cross-reactive responses, whereas Rev, Tat, and Vpu have even fewer conserved nine-mers than the highly variable Env protein, iii) M-group-optimized mosaic sets covered single subtypes nearly as well as within-subtype optimized sets, particularly for more conserved proteins.

Gag and the central region of Nef meet our four selection criteria. Nef is the HIV protein most frequently recognized by T-cells[21] and is the target for the earliest response in natural infection[22]. While it is variable overall (**Fig. 1e**), its central region is as conserved as Gag (**Fig. 1b**). Although mosaics could be designed to maximize the potential coverage of even the most variable proteins (**Fig. 3**), the prospects for global coverage are better for conserved proteins. Improved vaccine protection in macaques has been demonstrated by adding Rev, Tat, and Nef to a vaccine containing Gag, Pol, and Env[23], but this was in the context of homologous challenge, where variability was not an issue. The extreme variability of regulatory proteins in circulating virus populations may preclude cross-reactive responses. In terms of conservation, Pol, Gag (particularly p24) and the central region of Nef (HXB2 positions 65-149) are the most promising potential immunogens (**Fig. 1,3**). Pol, however, is infrequently recognized during natural infection[21], so we did not include it in this initial immunogen design. The conserved portion of Nef contains the most highly recognized peptides in HIV-1[21], but as a protein fragment, should not effect Nef's immune inhibitory functions (*e.g.* HLA class I down-regulation[24]). Both Gag and Nef are densely packed with well-characterized $CD8^+$ and $CD4^+$ T-cell epitopes, presented by many different HLA molecules (HIV Molecular Immunology Database); notably, Gag-specific $CD8^+$[25] and $CD4^+$[6] T-cell responses have been associated with low viral set points in infected individuals[25].

To examine potential effects of geographic variation and input sample size, we did a limited test using published C-subtype Gag sequences. We assembled three data-sets of comparable size (two South African sets[26] and one non-South-African set), generated mosaics independently on each set, and tested the resulting mosaics against all three sets. Nine-mer coverage was slightly better for identical training and test sets (77-79% 9/9 coverage). With different training and testing sets, results were essentially equivalent using the two different South African data sets (73-75%), or either South African set with the non-South-African C-subtype set (74-76%). Thus between- and within-country approximated within-clade coverage, and no advantage to a country-specific C subtype mosaic design was evident.

Designing mosaics for Gag and Nef and comparing vaccine strategies. To evaluate within- and between-subtype cross-reactivity for various vaccine design strategies, we calculated the coverage they provided for natural M-Group sequences. We computed the fraction of all perfect nine-mer matches between the natural sequences and the vaccine antigens, and (since single or double substitutions within epitopes may retain cross-reactivity) the proportion of 8/9 and 7/9 matches. **Figure 4** shows M group coverage per nine-mer in Gag and the central region of Nef for cocktails designed by various strategies: a) three non-optimal natural strains from the A, B, and C subtypes previously proposed as vaccine antigens[27]; b) three natural strains selected to give the best M group coverage; c) M group, B subtype, and C subtype consensus sequences; and, d,e,f) three, four and six mosaic proteins. For cocktails of *k* multiple strains, sets of *k=3*, *k=4*, and *k=6,* the mosaics clearly perform the best, and coverage approaches the upper bound for *k* strains. They are followed by optimally selected natural strains, the consensus protein cocktail, and finally, non-optimal natural strains. Allowing more antigens provides greater coverage, but gains for each addition are reduced as *k* increases (**Fig. 1,4**).

Figure 5 summarizes total coverage for the different vaccine design strategies, from single proteins through combinations of mosaic proteins, and compares within-subtype optimization to M group optimization. The performance of a single mosaic is comparable to the best single natural strain or a consensus sequence. Although a single consensus sequence out-performs the best single natural strain, the optimized natural-sequence cocktail does better than the consensus cocktail: the consensus sequences are more similar to each other than are natural strains, and are therefore somewhat redundant. Including even just two mosaic variants, however, markedly increases coverage, and four and six mosaic proteins give progressively better coverage than polyvalent cocktails of natural or consensus strains. Within-subtype optimized mosaics perform best – with four mosaic antigens 80-85% of the nine-mers are perfectly matched – but between-subtype coverage of these sets falls off dramatically, to 50-60%. In contrast, mosaic proteins optimized using the full M group give coverage of approximately 75-80% for individual subtypes, comparable to the coverage of the M group as a whole (**Fig. 5**, and **Supplementary Fig. S1**). If imperfect 8/9 matches are allowed, both M group optimized and within-subtype optimized mosaics approach 90% coverage. The C clade/B clade/M group comparisons presented (**Fig. 5)** are generally representative of within-clade, between-clade, and M group coverage, and coverage is not highly sensitive to input sequence representation. Despite a paucity of A and G clade sequences in our alignments, nine-mer coverage of A and G clade by

M-group optimized mosaics, though lower than coverage of B and C clade, was still high, and mosaic coverage of B and C clade was similar for both Gag and Nef, although there were more C-clade than B-clade Gag sequences, and more B-clade than C-clade Nef sequences in the input data (see **Supplementary Fig. S1** for a full comparison).

Since coverage is increased by adding progressively rarer nine-mers, and rare epitopes may be problematic (*e.g.*, by inducing vaccine-specific immunodominant responses), we investigated the frequency distribution of nine-mers in our vaccine constructs relative to the natural sequences from which they were generated. Most additional epitopes in a *k=6* cocktail compared to a *k=4* cocktail are low-frequency (<0.1, **Supplementary Fig. S2**). Despite enhancing coverage, these epitopes are relatively rare, so responses they induced might impair responses to more common, thus more useful, epitopes. Natural-sequence cocktails actually have fewer occurrences of moderately low-frequency epitopes than mosaics, which accumulate lower frequency nine-mers as coverage is optimized. However, our mosaics exclude unique and very rare nine-mers, which are present in nearly all natural strains. For example, natural M group Gag sequences had a median of 35 (range 0-148) unique nine-mers per sequence. We also explored retention of HLA-anchor motifs, and found anchor motif frequencies to be comparable between four mosaics and three natural strains. Natural antigens did exhibit an increase in number of motifs per antigen, possibly due to inclusion of strain-specific motifs (**Supplementary Fig. S3**).

The increase in ever-rarer epitopes with increasing cocktail size (*k*), coupled with concerns about antigen dilution and reagent development costs, led us to initially produce mosaic protein sets limited to 4 sequences (*k*=4), spanning Gag and the central region of Nef, optimized for subtype B, subtype C, and the M group (these sequences are included as supplementary data, as are mosaic sets for Env and Pol). Synthesis of various four-sequence Gag-Nef mosaics and initial antigenicity studies are underway (BHH, BH, NLL). Our initial mosaic vaccine, targets only Gag and the center of the Nef protein, which are conserved enough to provide excellent global population coverage, and have desirable properties (described above) in terms of natural responses[28]. Additionally, including B-subtype p24 variants in Elispot peptide mixtures to detect natural CTL responses to infection significantly enhanced both the number and the magnitude of responses detected (BK, KY, BDW and WF with Nicole Frahm and Christian Brander, manuscript in preparation), supporting the idea that including variants of even the most conserved proteins will be useful. Finally, cocktails of proteins in a polyvalent HIV-1 vaccine given to rhesus macaques did not interfere with the development of robust responses to each antigen[29], and antigen cocktails did not produce antagonistic responses in a mouse model[30], indicating that antigenic mixtures are appropriate for T-cell vaccines.

Even with mosaics, only limited nine-mer coverage is possible for variable proteins like Env, although mosaics improve coverage relative to natural strains. For example, three M group natural proteins (one each from the A, B, and C clades) currently under study for vaccine design[29] perfectly match only 39% of the nine-mers in M group proteins, and 65% have at least 8/9 matches. In contrast, three M group Env mosaics match 47% of nine-mers perfectly, and 70% have at least an 8/9 match. The code we have written to design polyvalent mosaic antigens is available (see Methods), and could readily be applied to any input set of variable proteins, optimized for any desired number of antigens. Our code also allows selection of optimal combinations of *k* natural strains, enabling rational selection of natural antigens for polyvalent vaccines, and we include in the supplement both mosaic sequences and the best natural strains for Gag and Nef population coverage of current database alignments.

Discussion

This study focuses on the design of T-cell vaccine components to counter HIV diversity at the moment of infection, and to block viral escape routes and thereby minimize disease progression in infected individuals. The polyvalent mosaic protein strategy developed here for HIV-1 vaccine design could be applied to any variable protein, to other pathogens, and to other immunological problems. For example, incorporating a minimal number of variant peptides into T-cell response assays could markedly increase sensitivity without excessive cost: a set of *k* mosaic proteins provides the maximum coverage possible for *k* antigens.

We previously proposed a centralized (consensus or ancestral) gene and protein strategy to address HIV diversity[3]. Proof-of-concept for the use of artificial genes as immunogens has been demonstrated by the induction of both T and B cell responses to wild-type HIV-1 strains by group M consensus immunogens[3,11-13]. The mosaic protein design improves on consensus or natural immunogen design by co-optimizing reagents for a polyclonal vaccine, excluding rare $CD8^+$ T-cell epitopes, and incorporating variants that, by virtue of their frequency at the population level, are likely to be involved in escape pathways.

The mosaic antigens maximize the number of epitope-length variants that are present in a small, practical number of vaccine antigens. We opted to use multiple antigens that resemble native proteins, rather than linking sets of concatenated epitopes in a poly-epitope pseudo-protein[31], reasoning that *in vivo* processing of native-like vaccine antigens will more closely resemble processing in natural infection, and will also allow expanded coverage of overlapping epitopes. T-cell mosaic antigens would be best employed in the context of a strong polyvalent immune response; improvements in other areas of vaccine design and a combination of the best strategies, incorporating mosaic antigens to cover diversity, may ultimately enable an effective cross-reactive vaccine-induced immune response against HIV-1.

Methods

HIV-1 sequence data. Reference alignments from the 2005 HIV sequence database were supplemented by recent C-subtype sequences from Durban, South Africa[26] for a worldwide sample of M-group sequences (551 Gag; 1,131 Nef) that included recombinants and pure-subtypes. Alignment subsets for within- and between-clade comparisons contained 18 A, 102 B, 228 C, and 6 G subtype sequences (Gag), and 62 A, 454 B, 284 C, and 13 G subtype sequences (Nef) .

The genetic algorithm. GAs apply computational analogues of biological evolution to problems that are difficult to solve analytically[32]. Solutions are evolved though random modification and selection according to a fitness (optimality) criterion. GAs come in many flavors; we implemented a “steady-state co-evolutionary multi-population” GA, in which candidate solutions are individually added to distinct populations that each contribute to the complete solution. A set of unaligned natural sequences is artificially recombined to generate a set of *k* pseudo-natural "mosaic" sequences, each containing sections of multiple natural sequences. Each population contributes one sequence to a cocktail, which is scored by *population coverage* [the proportion of all 9-amino-acid sequence fragments (potential epitopes) in the input sequences that are found in the cocktail].

To initialize the GA (**Fig. 2**), *k* populations of *n* initial candidate sequences are generated by 2-point recombination between randomly chosen natural sequences. Because the natural sequences are unaligned, crossover points are restricted to short strings (of length $c-1 = 8$, where a typical epitope length is $c = 9$) that match in both sequences. This ensures that the artificial recombinants resemble natural proteins: boundaries between sections of sequence from different strains are seamless, and local sequences are always found in nature. To prevent reduplication of repeats, the software explicitly prohibits excessive lengths. ("In frame" insertion of reduplicated epitopes could increase coverage without generating unnatural nine-mers, but would create "unnatural" proteins.) Initially, a cocktail contains one randomly chosen "winner" from each population. The fitness of any individual sequence is the coverage for a cocktail containing that sequence plus the current winners from the *other* populations, so the fitness of each sequence depends dynamically upon the winning sequences from the other populations.

Optimization proceeds one population at a time. For each iteration, two "parent" sequences are chosen. The first is the better of two sequences picked randomly from the current population ("2-tournament" selection). This selects parents with a probability inversely proportional to their fitness rank within the population, without having to compute all the fitnesses. The second parent is chosen the same way (half the time), or selected at random from the natural sequence input. A "child" sequence is generated by 2-point crossover between the parents. If the child contains any nine-mer found < 3 times in the natural population, it is rejected. Otherwise, it is scored and compared with four randomly chosen sequences from the same population. If any of those four sequences scores lower than the new (child) sequence, the child replaces it in the population. Whenever a child out-scores the current population "winner", it replaces the population winner in the cocktail. Typically 10 rounds of child generation are applied to each population in turn, cycling through the populations until evolution stalls (*i.e.*, no improvement has been made for a defined number of generations). The entire procedure is then restarted with new starting populations; restarts are repeated until no further improvement is seen. The GA was run on each data set with $n = 50$ or 500; each run was continued until no further improvement occurred for 12-24 hours on a 2 GHz Pentium processor. We generated cocktails having $k = 1, 3, 4,$ or 6 mosaic sequences.

Using our GA, one can optionally include one or more fixed sequences (*e.g.* a consensus) in the cocktail; other cocktail sequences will evolve to complement the fixed strain(s). An additional program selects from an input file the *k* intact natural strains that together provide the best population coverage.

Comparison with other polyvalent vaccine candidates. We compared population coverage for various potential mono- or polyvalent vaccines to the coverage of our mosaic-sequence vaccines, tracking exact nine-mer matches and 8/9 and 7/9 partial matches. nine-merPotential natural-strain vaccine candidates include single strains (*e.g.*, a single C strain for a South African vaccine[5]) or combinations of natural strains (*e.g.*, one each of subtypes A, B, and C[27]). To date, natural-strain vaccine candidates have not been selected to maximize T-cell epitope coverage; we picked vaccine candidates from the literature as plausible representatives of unselected natural-strain vaccines. We also determined an upper bound for coverage using only intact natural strains: for optimal natural-sequence cocktails, we selected the single sequence with the best coverage of the dataset, and successively added the best complements up to a given *k*. We scored optimal-natural-sequence cocktails of various sizes, as well as consensus sequences (alone or combined[3]), to represent centralized synthetic vaccines. Finally, we used the fixed-sequence GA option to generate and score consensus-plus-mosaic combinations, which were essentially equivalent to all-mosaic combinations for a given *k* (data not shown).

Additional Information. The code for these analyses is available at ftp://ftp-t10/pub/btk/mosaics. Reference alignments for this study were downloaded from the 2005 HIV sequence database (http://hiv.lanl.gov). CD8$^+$ and CD4$^+$ T-cell epitope maps from the HIV Molecular Immunology Database are at http://www.hiv.lanl.gov//content/immunology/maps/maps.html.

Acknowledgements

This work was funded through an internal directed research (DR) grant for vaccine design at Los Alamos National Laboratory: (WF, SP, TB, JT, BTK, KY, RF, BHH), NIH HIVRAD grant P01 AI061734-01: (BH, NLL, BHH, BTK, TB), NIH contract N01-15442: (BDW, BK), and NIH R21 AI055421-01-A1 (KY), and the CHAVI consortium (BH, BH, NL). The authors thank J. H. Fischer for helpful comments on the manuscript.

Figure 1. Upper bounds on epitope coverage of HIV-1 M group Gag, Nef, and Env proteins. The upper bound for population coverage of nine-mers for increasing numbers of variants is shown, for *k = 1–8* variants. A sliding window of length nine was applied across aligned sequences, moving down by one position. Different colors denote results for different numbers of sequences. At each window, the coverage given by the *k* most common nine-mers is plotted for Gag (**a,b**), Nef (**c,d**) and Env gp120 (**e,f**). Gaps inserted to maintain the alignment are treated as characters. The diminishing returns of adding more variants are evident, since, as *k* increases, increasingly rare forms are added. In (**a**), (**c**), and (**e**) the scores for each consecutive nine-mer are plotted in their natural order to show how diversity varies in different protein regions. In (**b**), (**d**) and (**f**) the scores for each nine-mer are reordered by coverage (a strategy also used in **Fig. 4**), to illustrate the spatial distribution of coverage for a given protein. Coverage is high for portions of Gag (e.g. P24; **a**) and the central region of Nef (**b**), and poor for gp120 (**c,f**).

Figure 2. Mosaic initialization, scoring, and optimization. (**a**) A set of *k* populations is generated by random 2-point recombination of natural sequences (we have tested 1-6 populations of 50-500 sequences each). One sequence from each population is chosen (initially at random) for the mosaic cocktail, which is subsequently optimized. The cocktail sequences are scored by computing *coverage* (defined as the mean fraction of natural-sequence nine-mers included in the cocktail, averaged over all natural sequences in the input data set). Any new sequence that covers more epitopes will increase the score of the whole cocktail. (**b**) The fitness score of any individual sequence is the coverage of a cocktail containing that sequence plus the current representatives from other populations. (**c**) Optimization: (**c.1**) two "parents" are chosen:

the higher-scoring of a randomly chosen pair of recombined sequences, and either (with 50% probability) the higher-scoring sequence of a second random pair, or a randomly chosen natural sequence. (**c.2**) Two-point recombination between the two parents is used to generate a "child" sequence. If the child contains unnatural or rare nine-mers it is immediately rejected; otherwise it is scored (**c.3**). If the score is higher than that of any of four randomly-selected population members, the child is inserted in the population in place of the weakest of the four, thus evolving an improved population; (**c.4**) if its score is a new high score, the new child replaces the current cocktail member from its population. Ten cycles of child generation are repeated for each population in turn, and the process iterates until improvement stalls.

Figure 3. Mosaic strain coverage for all HIV proteins. The level of nine-mer coverage achieved by sets of four mosaic proteins for each HIV protein is shown, with mosaics optimized using either the M group or the C subtype. The fraction of C subtype sequence nine-mers covered by mosaics optimized on the C subtype (within-clade optimization) is shown in gray. Coverage of nine-mers in non-C-subtype M-group sequences by C-subtype-optimized mosaics (between-clade coverage) is shown in white. Coverage of C-subtype sequences by M-group optimized mosaics is shown in black. B clade comparisons gave comparable results (data not shown).

Figure 4. Coverage of M group sequences by different vaccine candidates, nine-mer by nine-mer. Each plot presents site-by-site coverage (*i.e.,* for each nine-mer) of an M-group natural-sequence alignment by a single tri-valent vaccine candidate. Bars along the x-axis represent the proportion of sequences matched by the vaccine candidate for a given alignment position: 9/9 matches (in red), 8/9 (yellow), 7/9 (blue). Aligned nine-mers are sorted along the x-axis by exact-match coverage value. 656 positions include both the complete Gag and the central region of Nef. For each alignment position, the maximum possible matching value (*i.e.*the proportion of aligned sequences without gaps in that nine-mer) is shown in gray. (**a**) Non-optimal natural sequences selected from among strains being used in vaccine studies[27] including an individual clade A, B, and C viral sequences (Gag: GenBank accession numbers AF004885, K03455, and U52953; Nef core: AF069670, K02083, and U52953). (**b**) Optimum set of natural

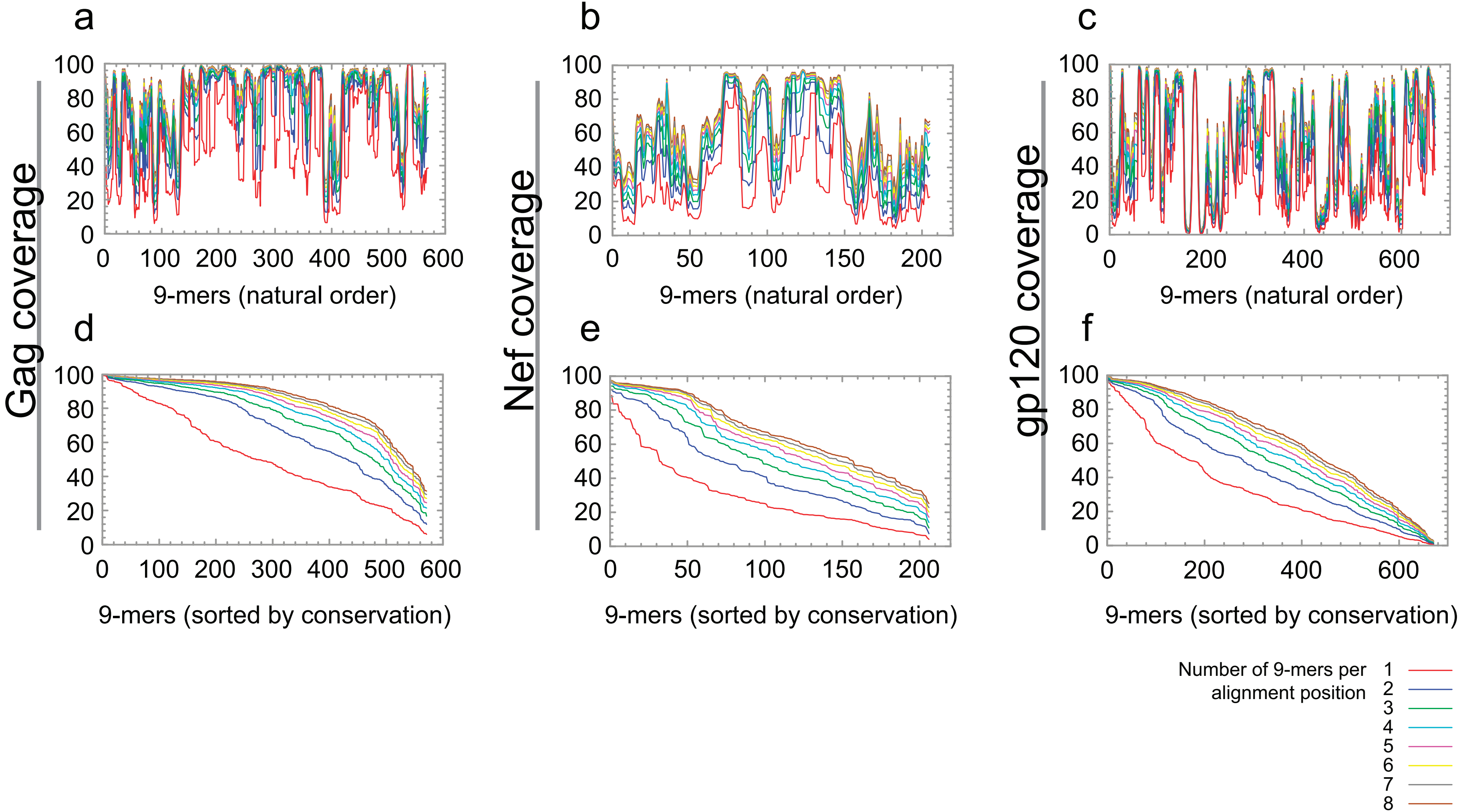
sequences [isolates US2 (subtype B, USA), 70177 (subtype C, India), and 99TH.R2399 (subtype CRF15_01B, Thailand); accession numbers AY173953, AF533131, and AF530576] selected by choosing the single sequence with maximum coverage, followed by the sequence that had the best coverage when combined with the first (i.e. the best complement), and so on, selected for M group coverage (**c**) Consensus sequence cocktail (M group, B- and C-subtypes). (**d**) 3 mosaic sequences, (**e**) 4 mosaic sequences, (**f**) 6 mosaic sequences. **d-f** were all optimized for M group coverage.

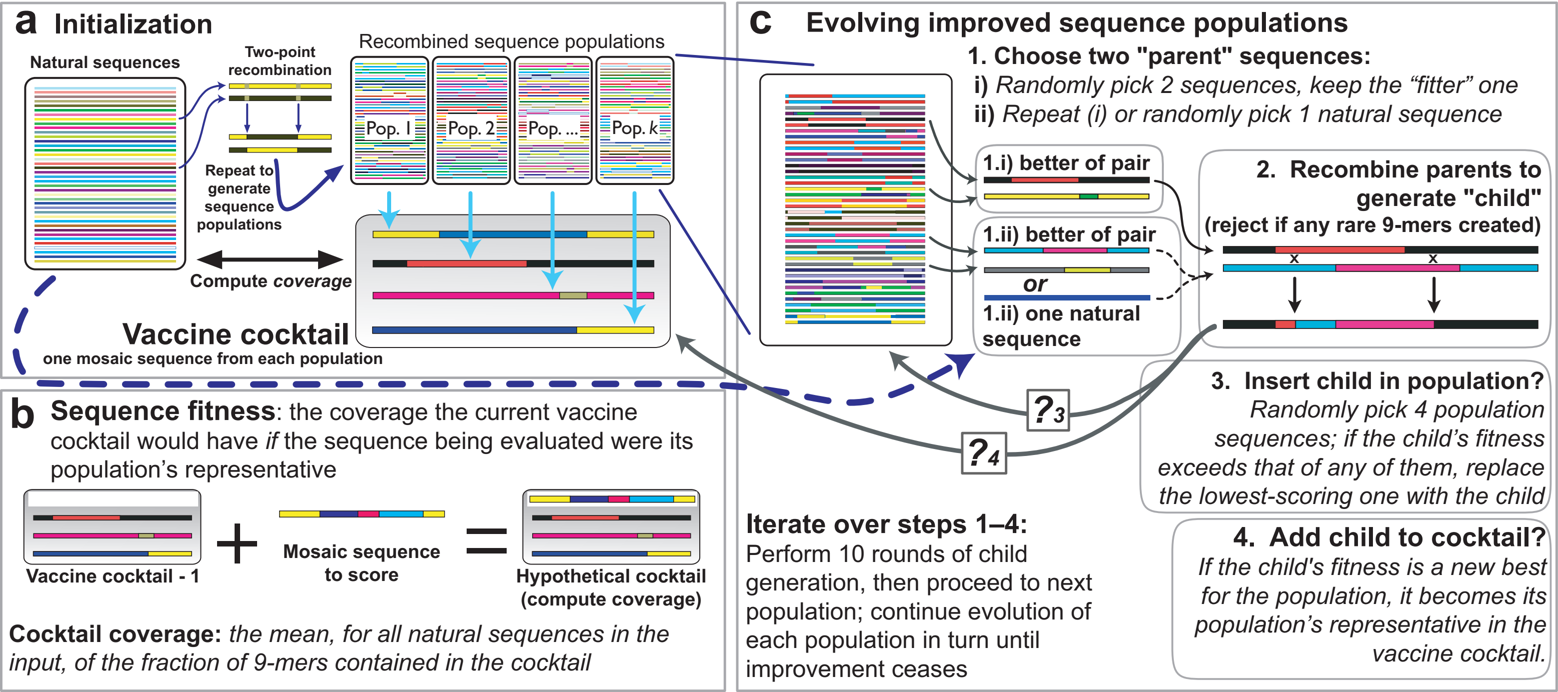
Figure 5. Overall coverage of vaccine candidates: coverage of nine-mers in C clade and M group sequences using different input data sets for mosaic optimization, allowing different numbers of antigens, and comparing different candidate vaccines. Exact (blue), 8/9 (one-off; red), and 7/9 (two-off; yellow) coverage was computed for mono- and polyvalent vaccine candidates for Gag and Nef (core) for four test situations: **within-clade** (C-clade-optimized candidates scored for C-clade coverage), **between-clade** (B-clade-optimized candidates scored for C-clade coverage), **global-against-single-subtype** (M-group-optimized candidates scored for C-clade coverage), **global-against-global** (M-group-optimized candidates scored for global coverage). Within each set of results, vaccine candidates are grouped by number of sequences in the cocktail (1-6); mosaic sequences are plotted with darker colors. “Non-opt” refers to one set of sequences moving into vaccine trials[27]; “mosaic” denotes sequences generated by our genetic algorithm; “opt. natural” denotes intact natural sequences selected for maximum nine-mer coverage; “MBC consensus” denotes a cocktail of 3 consensus sequences, for M-group, B-subtype, and C-subtype. A dashed line marks the coverage of a 4-sequence set of M-group-optimized mosaics (73.7–75.6%).

References

1. Nabel, G.J. HIV vaccine strategies. *Vaccine* **20**, 1945–1947 (2002).
2. Altfeld, M.*, et al.* HIV-1 superinfection despite broad CD8$^+$ T-cell responses containing replication of the primary virus. *Nature* **420**, 434–439 (2002).
3. Gaschen, B.*, et al.* Diversity considerations in HIV-1 vaccine selection. *Science* **296**, 2354–2360 (2002).
4. Korber, B.*, et al.* Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.* **58**, 19–42 (2001).
5. Williamson, C.*, et al.* Characterization and selection of HIV-1 subtype C isolates for use in vaccine development. *AIDS Res. Hum. Retroviruses* **19**, 133–144 (2003).
6. Oxenius, A.*, et al.* HIV-specific cellular immune response is inversely correlated with disease progression as defined by decline of CD4$^+$ T cells in relation to HIV RNA load. *J. Infect. Dis.* **189**, 1199–1208 (2004).
7. Barouch, D.H.*, et al.* Control of viremia and prevention of clinical AIDS in rhesus monkeys by cytokine-augmented DNA vaccination. *Science* **290**, 486–492 (2000).
8. Schmitz, J.E.*, et al.* Control of viremia in simian immunodeficiency virus infection by CD8$^+$ lymphocytes. *Science* **283**, 857–860 (1999).
9. Barouch, D.H.*, et al.* Viral escape from dominant simian immunodeficiency virus epitope-specific cytotoxic T lymphocytes in DNA-vaccinated rhesus monkeys. *J. Virol.* **77**, 7367–7375 (2003).
10. Moore, J.P. & Burton, D.R. Urgently needed: a filter for the HIV-1 vaccine pipeline. *Nat. Med.* **10**, 769–771 (2004).
11. Gao, F.*, et al.* Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein. *J. Virol.* **79**, 1154–1163 (2005).
12. Doria-Rose, N.A.*, et al.* Human immunodeficiency virus type 1 subtype B ancestral envelope protein is functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating subtype B envelope. *J. Virol.* **79**, 11214–11224 (2005).
13. Weaver, E.A.*, et al.* Cross-subtype T cell immune responses induced by an HIV-1 Group M consensus Env immunogen. *J. Virol.* (in press).
14. Altfeld, M.*, et al.* Enhanced detection of human immunodeficiency virus type 1-specific T-cell responses to highly variable regions by using peptides based on autologous virus sequences. *J. Virol.* **77**, 7330–7340 (2003).
15. Jones, N.A.*, et al.* Determinants of human immunodeficiency virus type 1 escape from the primary CD8$^+$ cytotoxic T lymphocyte response. *J. Exp. Med.* **200**, 1243–1256 (2004).
16. Allen, T.M.*, et al.* De novo generation of escape variant-specific CD8$^+$ T-cell responses following cytotoxic T-lymphocyte escape in chronic human immunodeficiency virus type 1 infection. *J. Virol.* **79**, 12952–12960 (2005).
17. Feeney, M.E.*, et al.* HIV-1 viral escape in infancy followed by emergence of a variant-specific CTL response. *J. Immunol.* **174**, 7524–7530 (2005).
18. Killian, M.S.*, et al.* Clonal breadth of the HIV-1-specific T-cell receptor repertoire in vivo as determined by subtractive analysis. *AIDS* **19**, 887–896 (2005).

19. Milicic, A., *et al.* CD8$^+$ T Cell Epitope-Flanking Mutations Disrupt Proteasomal Processing of HIV-1 Nef. *J. Immunol.* **175**, 4618–4626 (2005).
20. Ammaranond, P., *et al.* A new variant cytotoxic T lymphocyte escape mutation in HLA-B27-positive individuals infected with HIV type 1. *AIDS Res. Hum. Retroviruses* **21**, 395–397 (2005).
21. Frahm, N., *et al.* Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities. *J. Virol.* **78**, 2187–2200 (2004).
22. Lichterfeld, M., *et al.* HIV-1 Nef is preferentially recognized by CD8 T cells in primary HIV-1 infection despite a relatively high degree of genetic diversity. *AIDS* **18**, 1383–1392 (2004).
23. Hel, Z., *et al.* Improved vaccine protection from simian AIDS by the addition of nonstructural simian immunodeficiency virus genes. *J. Immunol.* **176**, 85–96 (2006).
24. Blagoveshchenskaya, A.D., Thomas, L., Feliciangeli, S.F., Hung, C.H. & Thomas, G. HIV-1 Nef downregulates MHC-I by a PACS-1- and PI3K-regulated ARF6 endocytic pathway. *Cell* **111**, 853–866 (2002).
25. Masemola, A., *et al.* Hierarchical targeting of subtype C human immunodeficiency virus type 1 proteins by CD8$^+$ T cells: correlation with viral load. *J. Virol.* **78**, 3233–3243 (2004).
26. Kiepiela, P., *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432**, 769–775 (2004).
27. Kong, W.P., *et al.* Immunogenicity of multiple gene and clade human immunodeficiency virus type 1 DNA vaccines. *J. Virol.* **77**, 12764–12772 (2003).
28. Bansal, A., *et al.* CD8 T-cell responses in early HIV-1 infection are skewed towards high entropy peptides. *AIDS* **19**, 241–250 (2005).
29. Seaman, M.S., *et al.* Multiclade human immunodeficiency virus type 1 envelope immunogens elicit broad cellular and humoral immunity in rhesus monkeys. *J. Virol.* **79**, 2956–2963 (2005).
30. Singh, R.A., Rodgers, J.R. & Barry, M.A. The role of T cell antagonism and original antigenic sin in genetic immunization. *J. Immunol.* **169**, 6779–6786 (2002).
31. Hanke, T., Schneider, J., Gilbert, S.C., Hill, A.V. & McMichael, A. DNA multi-CTL epitope vaccines for HIV and Plasmodium falciparum: immunogenicity in mice. *Vaccine* **16**, 426–435 (1998).
32. Holland, J.H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, (M.I.T. Press, Cambridge, MA, 1992).


a
b
c
Gag coverage
Nef coverage
gp120 coverage
100
80
60
40
20
0
0
100
200
300
400
500
600
50
150
9-mers (natural order)
d
e
f
9-mers (sorted by conservation)
Number of 9-mers per alignment position
1
2
3
4
5
6
7
8

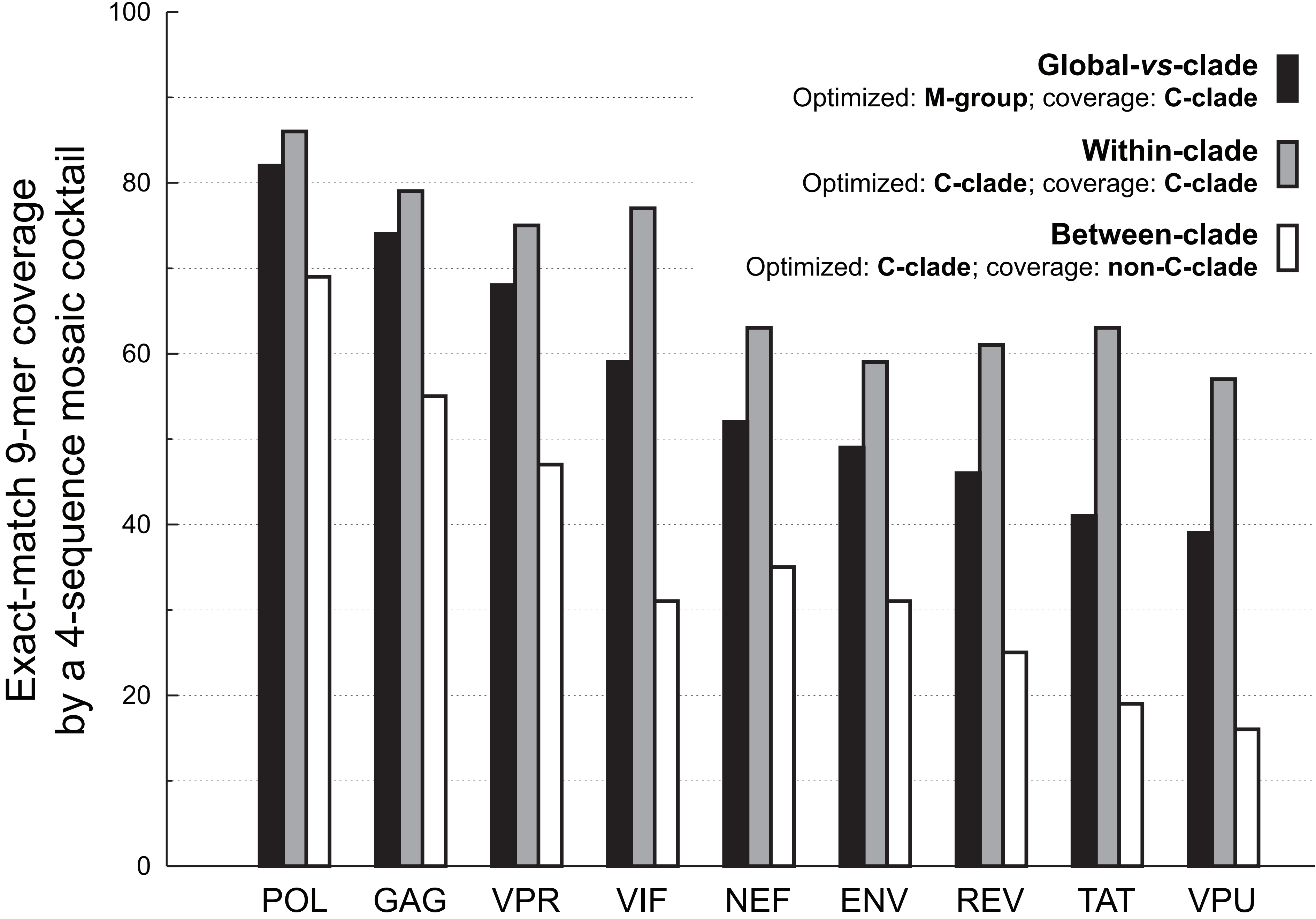

a
Initialization
Natural sequences
Two-point recombination
Repeat to generate sequence populations
Recombined sequence populations
Pop. 1
Pop. 2
Pop. ...
Pop. k
Compute *coverage*
Vaccine cocktail
one mosaic sequence from each population
b
Vaccine cocktail - 1
Mosaic sequence to score
Hypothetical cocktail (compute coverage)
c
Evolving improved sequence populations
1.i) better of pair
1.ii) better of pair
or
1.ii) one natural sequence
x
x
?3
?4

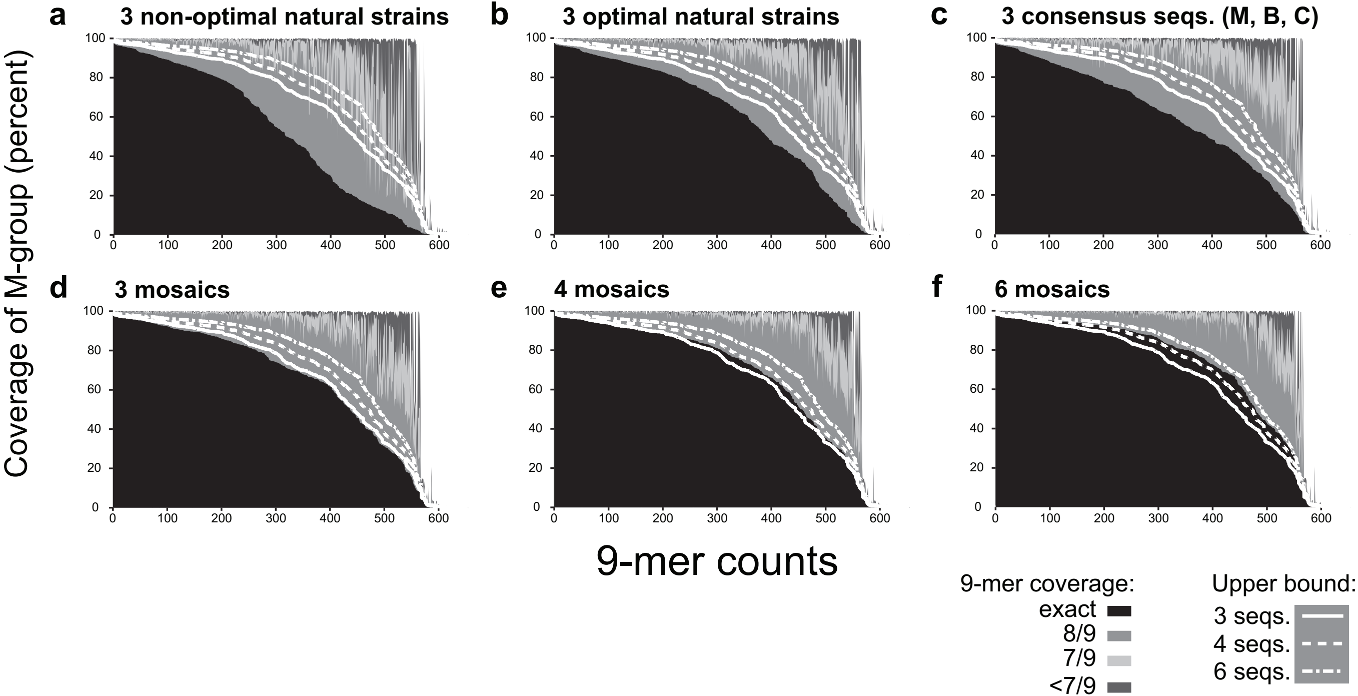
Sequence fitness: the coverage the current vaccine cocktail would have *if* the sequence being evaluated were its population's representative

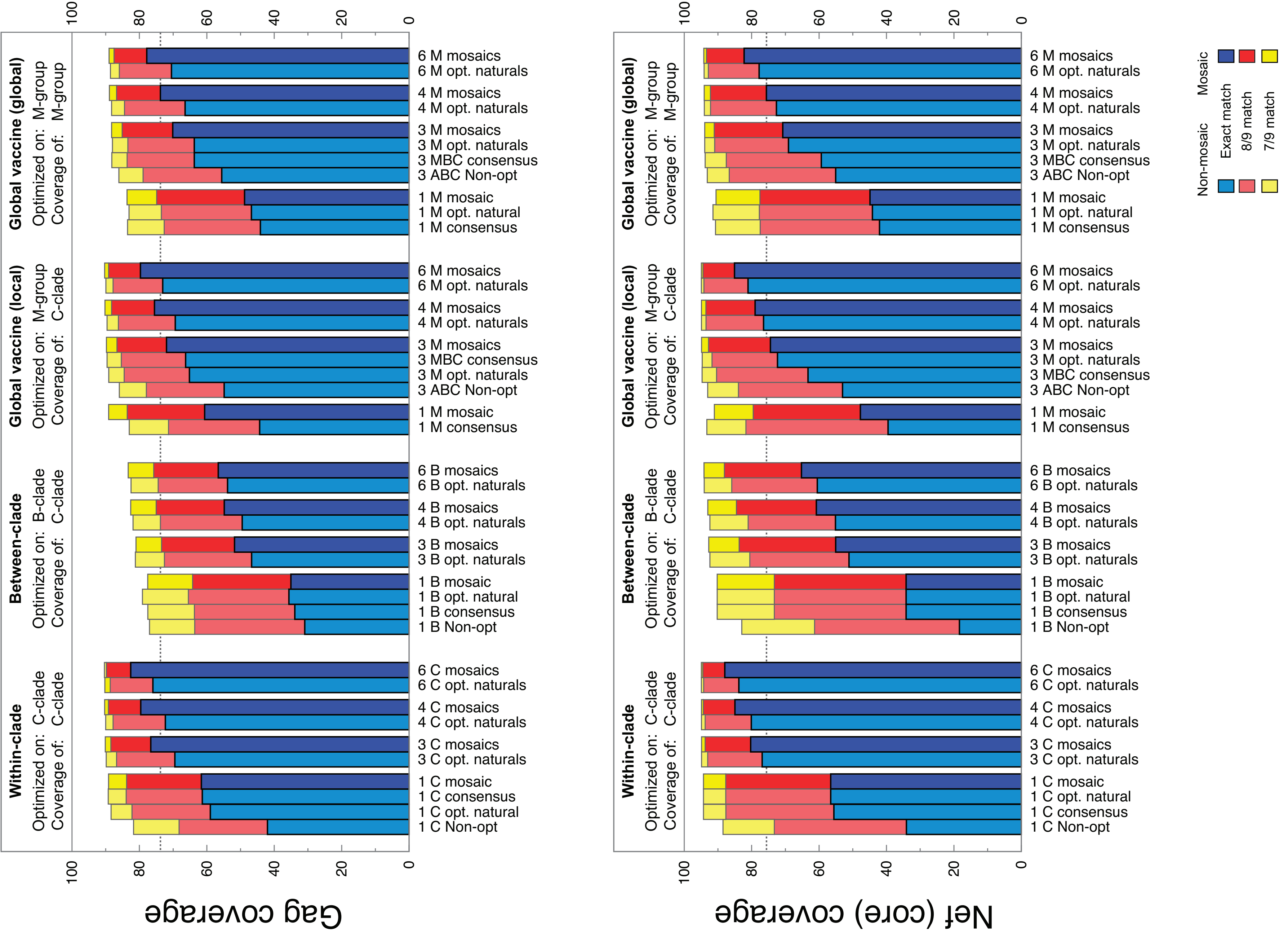
Cocktail coverage: *the mean, for all natural sequences in the input, of the fraction of 9-mers contained in the cocktail*

1. Choose two "parent" sequences:
   - i) *Randomly pick 2 sequences, keep the "fitter" one*
   - ii) *Repeat (i) or randomly pick 1 natural sequence*
2. Recombine parents to generate "child" (reject if any rare 9-mers created)
3. Insert child in population? *Randomly pick 4 population sequences; if the child's fitness exceeds that of any of them, replace the lowest-scoring one with the child*
4. Add child to cocktail? *If the child's fitness is a new best for the population, it becomes its population's representative in the vaccine cocktail.*

Iterate over steps 1–4:
Perform 10 rounds of child generation, then proceed to next population; continue evolution of each population in turn until improvement ceases


Exact-match 9-mer coverage
by a 4-sequence mosaic cocktail
100
80
60
40
20
0
POL
GAG
VPR
VIF
NEF
ENV
REV
TAT
VPU
Global-*vs*-clade
Optimized: **M-group**; coverage: **C-clade**
Within-clade
Optimized: **C-clade**; coverage: **C-clade**
Between-clade
Optimized: **C-clade**; coverage: **non-C-clade**


a
3 non-optimal natural strains
b
3 optimal natural strains
c
3 consensus seqs. (M, B, C)
d
3 mosaics
e
4 mosaics
f
6 mosaics
Coverage of M-group (percent)
0
20
40
60
80
100
0
100
200
300
400
500
600
9-mer counts
9-mer coverage:
exact
8/9
7/9
<7/9
Upper bound:
3 seqs.
4 seqs.
6 seqs.


Gag coverage
Net (core) coverage
0
20
40
60
80
100
Within-clade
Optimized on: C-clade
Coverage of: C-clade
Between-clade
Optimized on: B-clade
Coverage of: C-clade
Global vaccine (local)
Optimized on: M-group
Coverage of: C-clade
Global vaccine (global)
Optimized on: M-group
Coverage of: M-group
6 M mosaics
6 M opt. naturals
4 M mosaics
4 M opt. naturals
3 M mosaics
3 M opt. naturals
3 MBC consensus
3 ABC Non-opt
1 M mosaic
1 M opt. natural
1 M consensus
6 M mosaics
6 M opt. naturals
4 M mosaics
4 M opt. naturals
3 M mosaics
3 MBC consensus
3 M opt. naturals
3 ABC Non-opt
1 M mosaic
1 M consensus
6 B mosaics
6 B opt. naturals
4 B mosaics
4 B opt. naturals
3 B mosaics
3 B opt. naturals
1 B mosaic
1 B opt. natural
1 B consensus
1 B Non-opt
6 C mosaics
6 C opt. naturals
4 C mosaics
4 C opt. naturals
3 C mosaics
3 C opt. naturals
1 C mosaic
1 C consensus
1 C opt. natural
1 C Non-opt
6 M mosaics
6 M opt. naturals
4 M mosaics
4 M opt. naturals
3 M mosaics
3 M opt. naturals
3 MBC consensus
3 ABC Non-opt
1 M mosaic
1 M opt. natural
1 M consensus
6 M mosaics
6 M opt. naturals
4 M mosaics
4 M opt. naturals
3 M mosaics
3 M opt. naturals
3 MBC consensus
3 ABC Non-opt
1 M mosaic
1 M consensus
6 B mosaics
6 B opt. naturals
4 B mosaics
4 B opt. naturals
3 B mosaics
3 B opt. naturals
1 B mosaic
1 B opt. natural
1 B consensus
1 B Non-opt
6 C mosaics
6 C opt. naturals
4 C mosaics
4 C opt. naturals
3 C mosaics
3 C opt. naturals
1 C mosaic
1 C opt. natural
1 C consensus
1 C Non-opt
Non-mosaic
Mosaic
Exact match
8/9 match
7/9 match